二項ソフトクラスタリング分析例

この資料では、Visual Mining Studioのアイコン【Dyadic Soft Clustering】を使って、「二項ソフトクラスタリング」分析をする方法を説明します。二項ソフトクラスタリングは一般的にはPLSI, PLSAなどの名前で知られています。

株式会社NTTデータ数理システム

株式会社NTTデータ 数理システム

はじめに

NTT DATA Mathematical Systems Inc.

Visual Mining StudioのDyadic Soft Clusteringは次のようなデータの分析に適しています。

- ID付POSなど商品購買データ（トランザクションデータ）
- CookieIDのついた、Webページの閲覧記録（Webログ）
- 発言者IDと、発言ワードが対応付けされたデータ(典型的にはText Mining Studioの結果)

「誰が(ID)」「何を(商品コード)」「何個あるいは何回(数値)」を含むデータを対象としており、

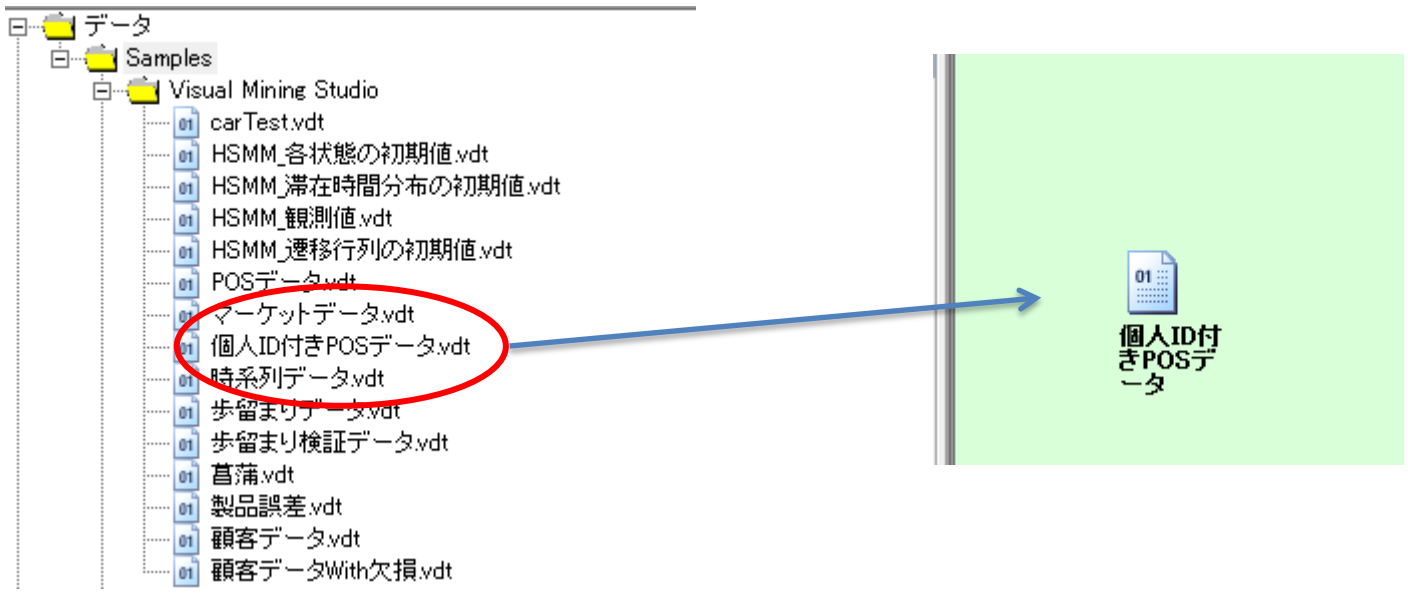
- リスト形式(縦持ちデータと呼ばれる）を対象にしているため、通常のクラスタ分析（k-means法など）のように、縦方向が「誰が」を表し、横方向が「何を」を表す横持ちと言われるデータを必要としません（横持ちデータはほとんどのセルがゼロになり、メモリ効率が非常に悪いデータです）
- 「誰が」をクラスタリングするだけでなく、「何を」もクラスタに分けることができます。商品購買データであれば、お客様のカテゴリと同時に、商品カテゴリの構築も可能です
- 「ソフトクラスタリング」はk-meansに代表される、「ハードクラスタリング」に対して、複数のクラスタに属することを許すクラスタリングを意味します

次ページから、Visual Mining Studioのサンプルデータ【個人ID付きPOSデータ】を例に、分析をご紹介します。

サンプルデータ取り込み

NTT DATA Mathematical Systems Inc.

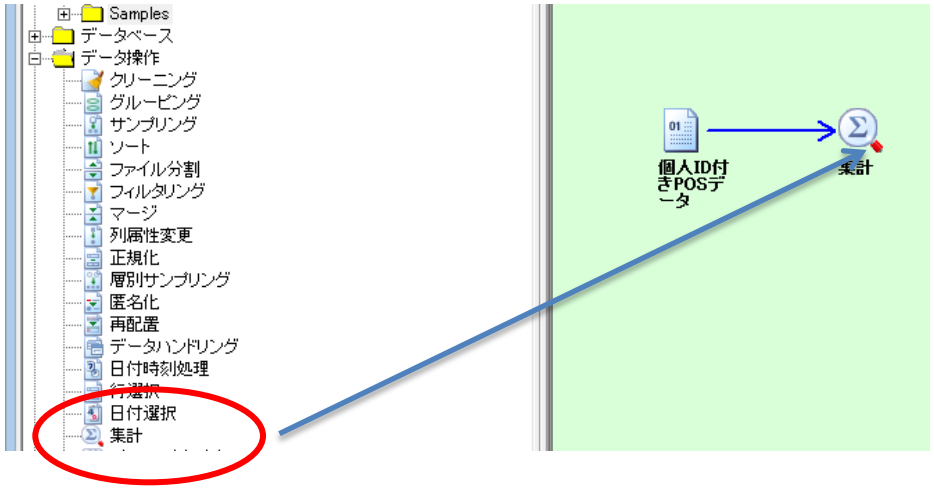
データはVisual Analytics Platform(VAP)の Object Browserから
【データ / Samples / Visual Mining Studio / 個人ID付きPOSデータ.vdt】
データを読み込み、用います。.vdtデータはVAP独自のデータ形式で、VAP上にはドラッグアンドドロップで張り付けて利用可能です。



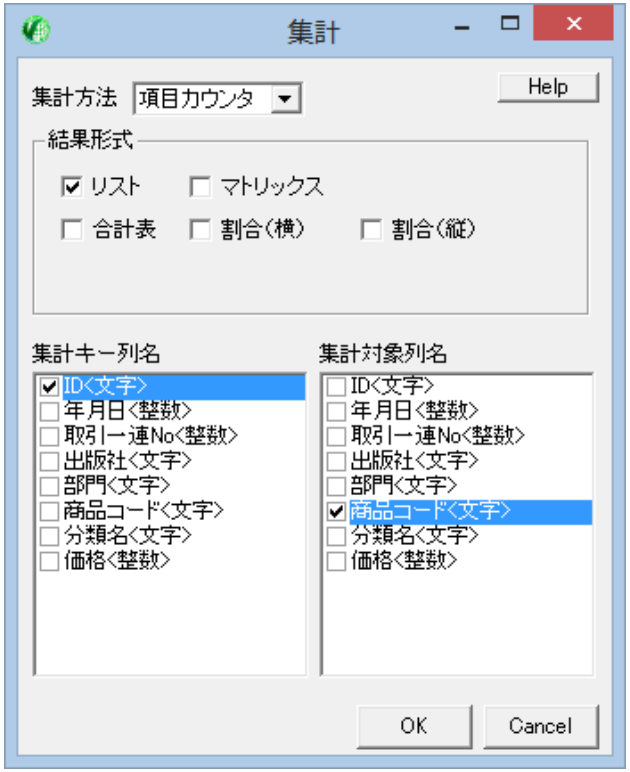
集計

NTT DATA Mathematical Systems Inc.

データには、ID(誰が)、商品コード(何を)買ったかが記録されています。二項ソフトクラスタリング分析にはこの2つと、重み（例として、何個買ったか、あるいは金額などのその購買の価値を図るための情報）が必要です。そこで、【集計】アイコンにより、ID と 商品コードのクロス集計をします。個数の情報がある場合は、集計キーを「IDと商品コード」、集計対象列を「個数」とし、個数の合計を計算してください。



【データ操作 / 集計】ドラッグアンドドロップ、アイコンのダブルクリック

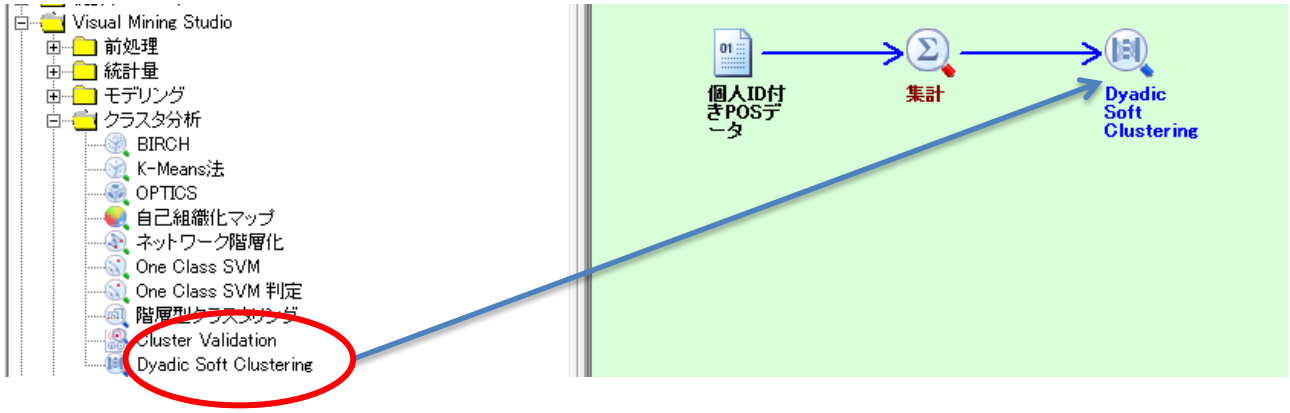


集計キー列名：ID
集計対象列名：商品コード
結果形式は必ず**リスト**にします

Dyadic Soft Clustering

NTT DATA Mathematical Systems Inc.

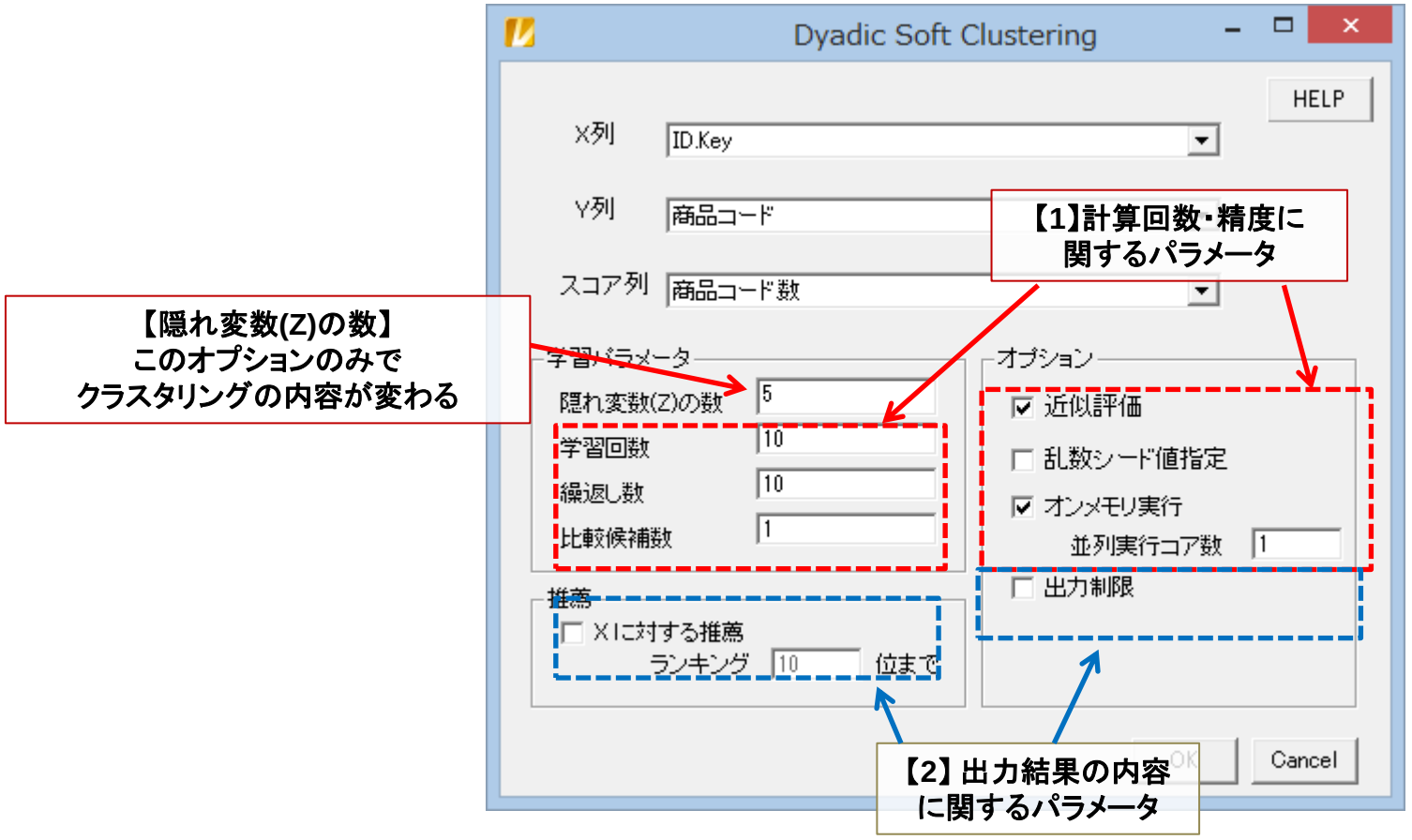
Dyadic Soft Clusteringアイコンをドラッグアンドドロップします。



Dyadic Soft Clustering

NTT DATA Mathematical Systems Inc.

X列には「誰が」の列を、**Y列には「何を」の列**を、**スコア列には「重み（何個、場合によっては金額でも）」の列**を指定します。また、【隠れ変数の数】には、想定しているクラスターの数を指定します。



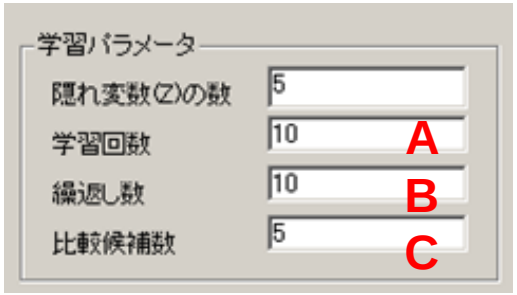
NTT DATA
NTT DATA Mathematical Systems Inc.

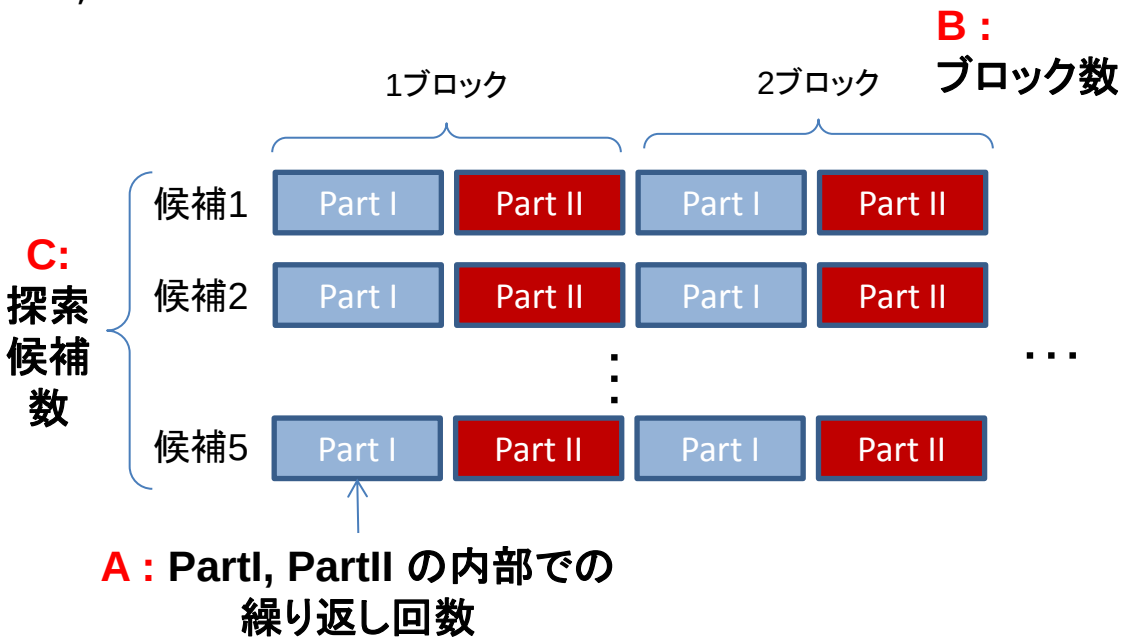
学習パラメータ

- お薦めのパラメータ設定
  - 学習回数 >= 10
  - 繰返し回数 >= 10
  - 比較候補数 >= 10
- 注意点
  - 繰返し回数 = 1 では発散するケースがあるので 2以上が必須です
  - お薦めパラメータ未満では、よりよい解が見つかるケースが多々あります（収束解とは程遠い値で止まってしまう）
    特に「学習回数」「繰返し回数」が10未満の場合は注意が必要であまりお勧めしません
  - データが大規模な場合、まず「比較候補数」を1～2として実行し、状況を確認した後に>=10とすることをお勧めします

学習パラメータ

- A:学習回数, B:繰返し回数, C:比較候補数





探索時間 は A×B×C に比例します。探索時間が長ければ通常は精度があがります。
A, B, C を偏りなく一定比率で増加させて、精度を上げるのがおすすめです。

結果を見る

NTT DATA
NTT DATA Mathematical Systems Inc.

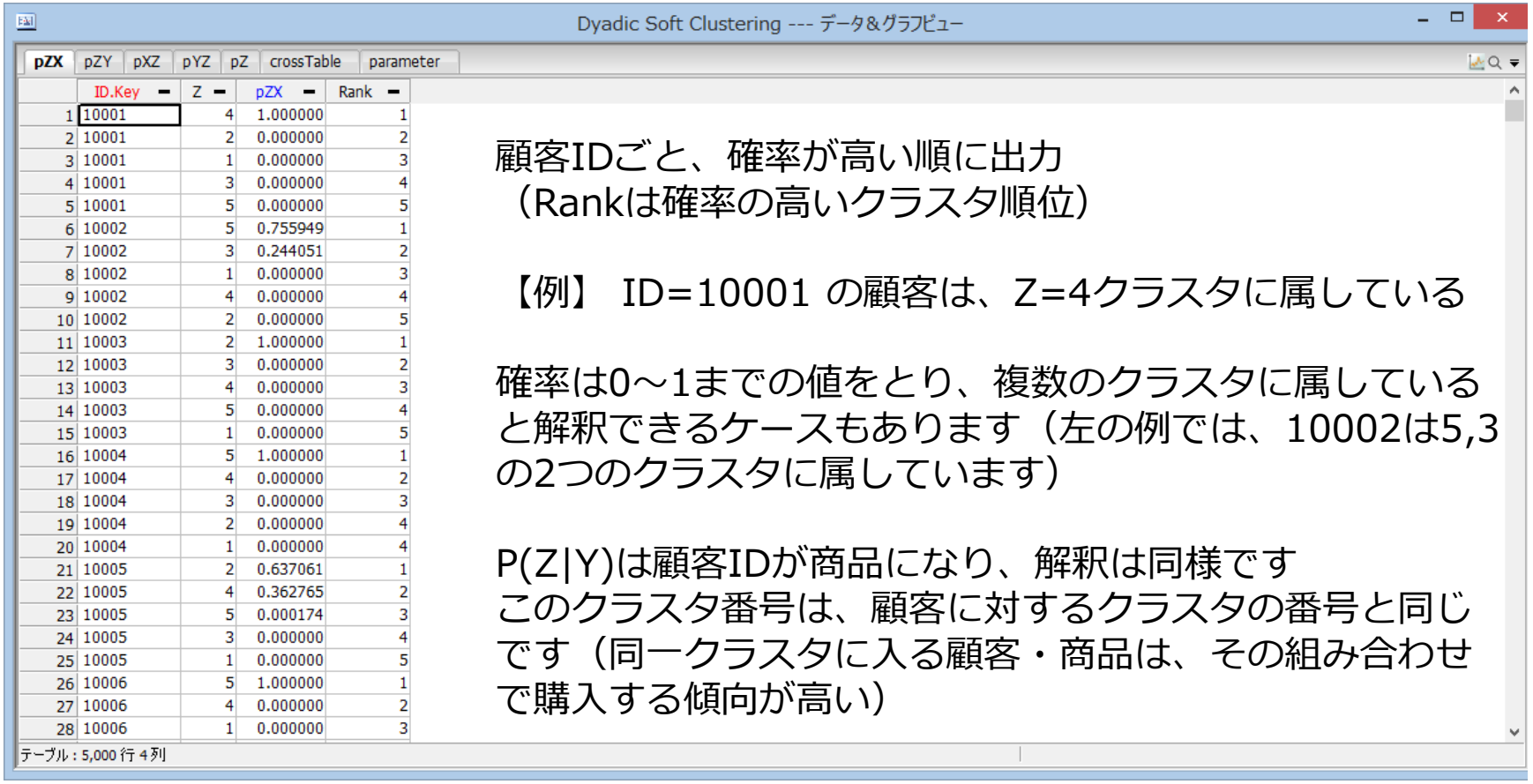
結果は複数のデータからなります
Xは「誰が」Yは「何を」Zは未知のクラスタを表します。Pはprobability(確率)を表します

データ名	内容
pZX	「誰が」が「どのクラスタ」に属するかを表す確率。人ごとに合計すると1になります。その人のクラスタ傾向を見るのに使います。
pZY	「何を」が「どのクラスタ」に属するかを表す確率。商品ごとに合計すると1になります。その商品のクラスタ傾向を見るのに使います。
pXZ	「誰が」が「どのクラスタ」に対する貢献度が高いかを見るのに使います。
pYZ	「何を」が「どのクラスタ」に対する貢献度が高いかを見るのに使います。
pZ	クラスタの出現確率を表します。
crossTable	クラスタごとの「誰が」x「何を」をクロス表にしたものです。対角要素が大きいことを確認して、クラスタの妥当性をみます。
parameter	各種の統計量などを表示します。隠れ変数(クラスタ）を変えて計算したときに、クラスタ数は何個が適切かどうか確認するのに使います。

結果の見方(1)

NTT DATA
NTT DATA Mathematical Systems Inc.

P(Z|X) … 顧客Xが、クラスタZに所属する確率
P(Z|Y) … 商品Yが、クラスタZに所属する確率



	ID.Key	Z	pZX	Rank
1	10001	4	1.000000	1
2	10001	2	0.000000	2
3	10001	1	0.000000	3
4	10001	3	0.000000	4
5	10001	5	0.000000	5
6	10002	5	0.755949	1
7	10002	3	0.244051	2
8	10002	1	0.000000	3
9	10002	4	0.000000	4
10	10002	2	0.000000	5
11	10003	2	1.000000	1
12	10003	3	0.000000	2
13	10003	4	0.000000	3
14	10003	5	0.000000	4
15	10003	1	0.000000	5
16	10004	5	1.000000	1
17	10004	4	0.000000	2
18	10004	3	0.000000	3
19	10004	2	0.000000	4
20	10004	1	0.000000	4
21	10005	2	0.637061	1
22	10005	4	0.362765	2
23	10005	5	0.000174	3
24	10005	3	0.000000	4
25	10005	1	0.000000	5
26	10006	5	1.000000	1
27	10006	4	0.000000	2
28	10006	1	0.000000	3

顧客IDごと、確率が高い順に出力
（Rankは確率の高いクラスタ順位）

【例】 ID=10001 の顧客は、Z=4クラスタに属している

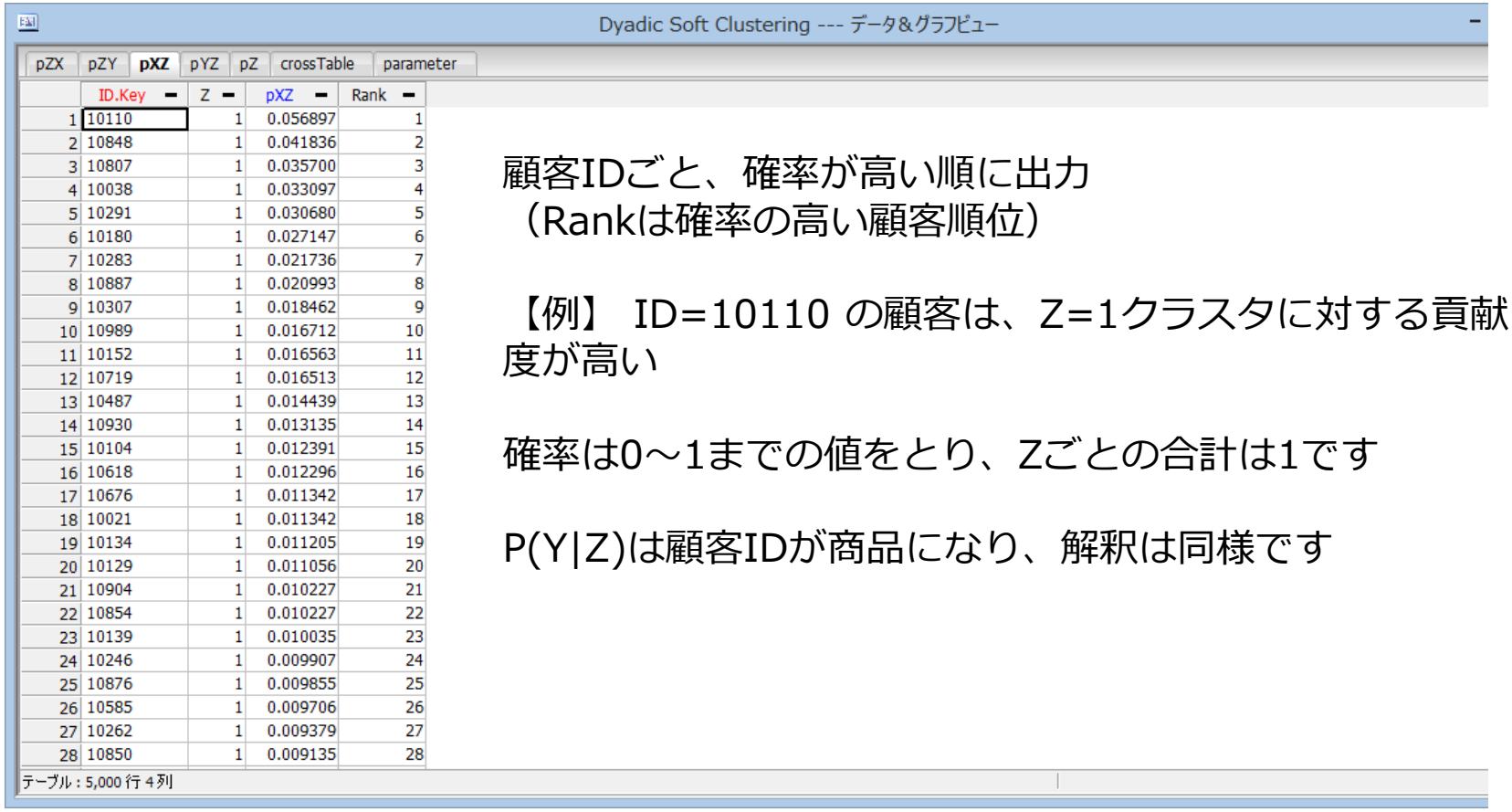
確率は0～1までの値をとり、複数のクラスタに属していると解釈できるケースもあります（左の例では、10002は5,3の2つのクラスタに属しています）

P(Z|Y)は顧客IDが商品になり、解釈は同様です
このクラスタ番号は、顧客に対するクラスタの番号と同じです（同一クラスタに入る顧客・商品は、その組み合わせで購入する傾向が高い）

結果の見方(2)

NTT DATA Mathematical Systems Inc.

P(X|Z) … 顧客Xのクラスタ Z内での貢献度を表す確率
P(Y|Z) … 商品Yのクラスタ Z内での貢献度を表す確率



	ID.Key	Z	pXZ	Rank
1	10110	1	0.056897	1
2	10848	1	0.041836	2
3	10807	1	0.035700	3
4	10038	1	0.033097	4
5	10291	1	0.030680	5
6	10180	1	0.027147	6
7	10283	1	0.021736	7
8	10887	1	0.020993	8
9	10307	1	0.018462	9
10	10989	1	0.016712	10
11	10152	1	0.016563	11
12	10719	1	0.016513	12
13	10487	1	0.014439	13
14	10930	1	0.013135	14
15	10104	1	0.012391	15
16	10618	1	0.012296	16
17	10676	1	0.011342	17
18	10021	1	0.011342	18
19	10134	1	0.011205	19
20	10129	1	0.011056	20
21	10904	1	0.010227	21
22	10854	1	0.010227	22
23	10139	1	0.010035	23
24	10246	1	0.009907	24
25	10876	1	0.009855	25
26	10585	1	0.009706	26
27	10262	1	0.009379	27
28	10850	1	0.009135	28

テーブル：5,000 行 4 列

顧客IDごと、確率が高い順に出力
（Rankは確率の高い顧客順位）

【例】 ID=10110 の顧客は、Z=1クラスタに対する貢献度が高い

確率は0～1までの値をとり、Zごとの合計は1です

P(Y|Z)は顧客IDが商品になり、解釈は同様です

2項クラスタリング –計算方法

NTT DATA Mathematical Systems Inc.

顧客・商品ごとの購入点数行列に対して、顧客・商品を入れ替えて、同時に買われている組をクラスタとして抽出します

顧客＼商品	1	2	3	4	5	6
A	1	0	0	2	0	0
B	0	0	1	0	4	2
C	0	1	0	0	3	0
D	2	0	0	1	0	0
E	0	0	0	0	5	0
F	0	0	1	0	0	1
G	0	0	0	1	0	0

クラスタリング後

顧客＼商品	1	4	2	5	3	6
A	1	2	0	0	0	0
D	2	1	0	0	0	0
G	0	1	0	0	0	0
C	0	0	1	3	0	0
E	0	0	0	5	0	0
B	0	0	0	4	1	2
F	0	0	0	0	1	1

クラスタ1

クラスタ2

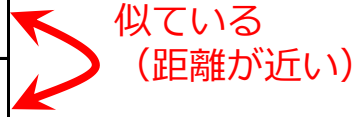
クラスタ3

(参考)通常よく使われているクラスタリングの計算方法 (k-means, 階層型クラスタリングなど)

NTT DATA Mathematical Systems Inc.

顧客・商品の購入行列に対して、顧客の行と行の距離を計算し、距離の近い顧客同士を同一クラスタに割り当てます。距離計算の方法には、ユークリッド距離、cosine距離、Manhattan 距離などがあります。

顧客＼商品	1	2	3	4	5
A	12000	5200	210	0	0
B	13000	4900	240	0	0
C	0	2420	15000	0	
D	0	12000	0	15000	0



2つのクラスタリングの違い

NTT DATA
NTT DATA Mathematical Systems Inc.

	2項クラスタリング	K-means法などのクラスタリング
入力データ	リストデータ（マトリックスデータの疎表現） X, Y, 購買個数のレコード並び。対応する組み合わせがない場合は、レコードそのものが出現しないので、少ないメモリでデータ記録が可能	マトリックスデータ 行:X, 列:Yとしてデータを表現。対応する組み合わせが出現しない場合、対応するセルを0とする。あまり買われない商品にも0と記録する必要があるので、メモリ量が多い
クラスタリング方法	顧客・商品の共起に基づく方法	顧客(X:行)の間の距離計算に基づく方法
クラスタリング結果	ソフトクラスタリング クラスタへの所属確率が0～1の間に決まる	ハードクラスタリング クラスタは1つのみに決まる
クラスタリング対象	顧客・商品の同時クラスタリング	顧客に対するクラスタリング

2項クラスタリング – 実運用上の注意点

- 値のスケール、範囲
  - 値の差が小さくなるようなパラメータが計算されるため、【スコア列】のスケールが重要です
  - POSデータの場合、商品の買い合わせ(同じバスケット）に着目して2項クラスタリングを実施するのが適切です。そのため、次の変数を使うのが適当です
    - 購入点数（金額は商品ごとの差が大きいので、あまり適しません）
    - 購入経験有無（買われたら1（買われなかったらデータなし）, 今回のサンプルプロジェクトはこちらのやり方です）
  - 現バージョンでは、ゼロ、あるいはマイナスの値を持つデータがあると正しく計算されませんので、【データハンドリング】などでデータをフィルタリングして利用してください

- POSデータ以外での活用
  - Cookie IDがついたWebページの閲覧ログ(IDとWebページのクラスタリング)
  - IDとタグの情報、ECサイトなどでの商品リストにタグがついているようなデータにも利用可能です
  - IDと発言された単語の組み合わせデータ(Text Mining Studioとの組み合わせ)

サンプルデータでお試しいただく二項ソフトクラスタリングはいかがでしたか？ぜひ、ご自身のデータでお試しください。また、分析詳細や各設定について、詳細はマニュアルをご覧ください。

保守ご契約中の方、テスト使用中の方は技術サポートサービスをご利用いただけます。
技術サポートはメールにて承っております。

※分析に関するご相談、あるいはプログラミングは技術サポートでは承っておりません。また、お電話でのお問い合わせには回答しておりませんので、ご了承ください。

【E-mail】 vmstudio-support@msi.co.jp
【URL】 http://www.msi.co.jp/vmstudio/

ライセンス、料金、その他製品に関するお問い合わせは下記NTTデータ数理システム営業部までお問い合わせください。

TEL：03 - 3358 – 6681
FAX：03 - 3358 – 1727

NTT DATA
株式会社NTTデータ 数理システム